Making the Standard More Standard: A Data and Query Model for Knowledge Representation in the Arden Syntax

Robert A. Jenders, MD, MS[1]; Roger Corman, BS[2]; Balendu Dasgupta, MS[1]

[1]Cedars-Sinai Medical Center and
Department of Medicine, University of California, Los Angeles
[2]Eclipsys Corporation, Santa Rosa, CA

Context: Arden Syntax is a Health Level Seven (HL7) standard that can be used to encode computable knowledge. However, dissemination of knowledge is hampered by lack of standard database linkages in Arden knowledge bases (KB). Moreover, the HL7 Reference Information Model (RIM) is object-oriented and hence incompatible with the current Arden data model. Also, significant investment has been made in Arden KBs that would be lost if a backward-incompatible data model were adopted. Objective: To define a data model that standardizes database linkages and provides object-oriented features while maintaining backward compatibility. Analysis: We identified the objects of the RIM that could be used as a schema for standard database queries. We propose extensions to Arden to accommodate this model, including the manipulation of objects. Conclusion: A data model that standardizes database linkages and introduces object-oriented constructs will facilitate knowledge transfer without violation of backward compatibility in the Arden Syntax.

INTRODUCTION

Challenge of Knowledge Sharing

Considerable delay often occurs between confirmation of a clinically relevant research finding in the medical literature and the incorporation of that finding into widespread clinical practice [1]. Barriers to the use of such knowledge include lack of awareness, lack of familiarity and inertia of previous practice [2]. In part to overcome such barriers to knowledge dissemination, clinical decision support systems (CDSS) have been recommended in order to improve patient safety [3]. Indeed, computer-based guideline implementation systems have been shown to improve both clinician performance and clinical outcomes [4].

Knowledge Base Sharing and Standards

Despite such success, use of computable knowledge at institutions other than those of initial development is limited. In part this is the result of the use of knowledge representation formalisms that are difficult to share with other institutions. Arden Syntax is an American National Standards Institute (ANSI) formalism supervised by Health Level Seven (HL7) for representation of procedural medical knowledge [5]. The unit of representation in the Syntax is the Medical Logic Module (MLM), which contains enough data and logic to make a single medical decision. The Arden Syntax has been implemented at multiple sites worldwide and is included in the software of several major vendors. The current version of Arden is 2.1, accepted by ANSI in December, 2002.

Admittedly, even with such a standard, some site-specific changes must occur in order for a knowledge base to be transferred from one site to another [6]. Key to minimizing site-specific changes is the standardization of database linkages, which in turn requires identification of a standard data model, vocabularies and query syntax [7]. Without such standardization, queries would have to be rewritten at each institution in order to match the local database schema, vocabularies and query language. This is sometimes known as the "curly braces problem" of Arden because of the syntactic construct used to enclose these site-specific references [8]. These constructs apply not only to database queries but also to declaration of triggering events and definition of the destination of CDSS output.

One such standard data model is the HL7 Reference Information Model (RIM), which is an object-oriented model that purports to describe all of the data that might be transmitted between health care computing systems in compliance with version 3 of the HL7 messaging standard. We previously analyzed the RIM as it existed in 1997 to represent queries in a typical knowledge base, and we found that it adequately represented most data elements contained in the queries [8].

However, the RIM has changed significantly since then. Moreover, use of an object-oriented data model to encode queries raises the challenge of how the

knowledge representation language can manipulate data returned from such queries. The data model of Arden Syntax is relatively simple; each variable resulting from a query has only two fixed attributes: value and primary time [9]. Moreover, operators that process such variables assume only this low level of complexity.

Work by others has used elements of the Arden Syntax to create an object-oriented expression language, including methods, that could be used in a formalism such as the GuideLine Interchange Format (GLIF) [10]. The expression language would be employed to denote the logic and database queries of a unit of knowledge. Other workers have demonstrated a specific niche for Arden in distinction to a guideline formalism, suggesting that these two formalisms would play complementary roles [11] and that Arden should not be subsumed by the guideline formalism.

However, inclusion of this expression language in the Arden Syntax would make it backward-incompatible with older versions of the Syntax. This, in turn, has raised concern among vendors and other organizations that have made significant investments in knowledge bases encoded in Arden. Using the most up-to-date version of Arden would require rewriting legacy MLMs to conform to the new standard, resulting in significant cost and possibly introducing new errors into an already validated knowledge base.

To help address the issue of standardization of database linkages, we previously developed a knowledge editor for the Arden Syntax [1]. However, this employed only an ad hoc approach to this issue by suggesting a query format without extending Arden to accommodate the resulting data model.

Goals of the Analysis

Accordingly, we have undertaken the present work with two important goals. The first is to provide a standard for encoding database linkages in the Arden Syntax. The second is to adapt the Syntax itself to provide limited object-oriented functionality that can process the results of such queries without eliminating backward compatibility.

BACKGROUND

Current Arden Database Linkages

The developers of the Arden Syntax in 1989 recognized that defining a standard data model, vocabularies and query syntax that could accommodate widely varying information system architectures would be a difficult challenge [9]. Accordingly, they deliberately omitted these items from the standard. Instead, local linkages are encoded inside curly braces, thus identifying them to compilers as well as to those institutions with whom the MLMs might be shared.

Nonetheless, part of a database query may be expressed using the standard. Aggregation operators and constraints that restrict the values returned by a query already are a part of the standard. These aggregation operators include *last*, *first*, *earliest*, *latest* and others. Constraints include restrictions based on the time and the value of the data. The general form of a database linkage is represented in the Arden READ statement. This is *<variable> := READ <aggregation> <mapping> WHERE <constraint>*. It is the mapping, then, which currently is not standardized and is the focus of the present work.

Object-Oriented Data in the Current Arden

The result of any query executed in an Arden MLM is a list of scalar values (e.g., strings, numbers, Boolean values as primitive types), each of which is a simple object with two, fixed attributes: value and primary time.

However, medical data may be naturally aggregated as objects that are more complex than this. For example, a blood pressure may have a number of attributes—measurement time, systolic pressure, position of patient and so on--that may be logically linked in a database as different attributes of an object. Nevertheless, the only way in the current version of Arden to manipulate these results is as parallel lists. In this way, each attribute of an object is represented as a distinct list, and the nth element of each list pertains to the same measurement. Thus, the burden falls on the programmer to keep track of these lists and ensure that their elements remain parallel.

Each operator in the current Arden is defined only with regard to these data primitives. For example, the addition operator has a defined result if its operands are either numbers or lists of numbers but is otherwise undefined. Thus, Arden operators would have to be redefined in order to manipulate arbitrarily complex objects as operands.

Vocabularies in the Current Arden

Unfortunately, no widespread agreement on standards in this area exists. Nonetheless, candidate vocabularies exist for particular disciplines, such as LOINC for representing concepts concerning laboratory results [12]. Another candidate vocabulary to describe elements such as problem lists, allergies and related items is SNOMED CT [13].

Typical queries in current installation of Arden include a wide range of vocabularies. These range from these standards, to vendor-specific vocabularies to local, enterprise-wide vocabularies such as the Medical Entities Dictionary of Columbia-Presbyterian Medical Center [14]. References to these vocabularies commonly are part of queries encoded in the Structured Query Language (SQL), but any number of query syntaxes—such as parameters for a query optimizer or data access modules—exist [5].

PROPOSED NEW ARDEN DATA MODEL

Query Model

In light of the ubiquity of relational clinical data repositories, the proposed new data model for Arden includes the use of SQL as the query syntax. In light of Arden as a standard in the HL7 suite of standards, the proposed model uses the RIM to define the objects against which a query would be executed. References to the RIM use version 1.21.

The general form is

<variable> :=
READ <aggregation> <attribute>
FROM <RIM object>
WHERE <constraint>;

In this form, <aggregation> and <constraint> are the same as in the standard part of the current Arden. The <object> is the relevant element in the RIM, and the <attribute> is a property of that element. The <attribute> may be a list of properties, in which case <variable> should be a list of variables, one for each object attribute. The object included in the query would be that object from the RIM that either directly possesses or inherits the desired attributes.

When a constraint involves a reference to a standard vocabulary, we propose that this reference be encoded using a triplet:

*'<code>'^'<code name>'^*
*'<vocabulary name>'^*
'<vocabulary version>'

This similar notation may be used when defining a trigger event in Arden, which is another kind of institution-specific mapping that identifies the circumstance when the MLM should be executed.

Using the RIM, we identify several classes of queries or templates based on the kind of data being retrieved. One such class is demographic queries. In the RIM hierarchy, a person (having attributes such as addr) is a living subject (having attributes such as administrativeGenderCode) which in turn is an entity (having attributes such as name). An example of a demographic query under this model is *(name, sex, location) := read name, administrativeGenderCode, addr from person where name = 'Jones'.*

Another query archetype involves observations, such as laboratory test results or diagnoses. In the RIM, an observation (having attributes such as value and interpretationCode) is an act (having attributes such as code, classCode and moodCode). To define an observation, we set classCode to 'OBS' and moodCode to 'EVN' (for an actual event). An example of a query under this model for a laboratory result encoded using LOINC is *plasma_cell_count := read value from observation where code='24103-4'^'PLASMA CELLS'^'LN'^'2.05' and classCode = 'OBS' and moodCode='EVN'.*

A third key query archetype involves medication. In this situation, the class substanceAdministration (having attributes such as routeCode) is an act (having attributes as above). To define a medication record, we set classCode to 'SBADM' and moodCode to 'EVN'. An example of a query to retrieve the coded names of all a patient's oral medication is *oral_meds := read code from substanceAdministration where routeCode = 'PO' and classCode = 'SBADM' and moodCode = 'EVN'.*

Other query archetypes include clinical encounters and the results of past CDSS activities, e.g., alerts.

Processing Objects

Having included an object-oriented data model as part of the query structure, we must address how to handle the data that are returned by such queries and stored in variables in the MLM. Provided that a query reads specific attributes of objects in the database and those attributes are current primitive data types, no further extension of Arden is necessary. However, introducing objects into Arden will eliminate the challenge of maintaining parallel lists to group related data and thus will ease the process of retrieving those data from the clinical database.

To do this, we introduce an object statement definition. Its form is

<variable> := OBJECT [<attribute-1>, <attribute-2>,...]

where each attribute is a valid Arden identifier. In this way, logically related data can be grouped into the same variable. We further introduce a "dot notation" in order to access individual attributes of an object. In this way, a reference to *object.attribute* will yield the value of that attribute. Each attribute would have the type of a current Arden primitive or would

itself be an object. In the latter case, the dot notation would be composed to reference nested attributes of an object, e.g. *object.attribute.attribute*.

In order to facilitate direct retrieval of objects from a database, we propose extension of the current READ statement by adding a READ AS statement which serves as a structured READ. The general form of this new statement is

<variable> := READ AS <object type>
<aggregation> (<mapping>) WHERE
<constraint>;

Instead of returning a list of attributes from a database relation, this query returns a list of rows or objects. Individual attributes in an object variable could be assigned either in order of occurrence in the declaration or by name (if an exact match with a RIM attribute name is used).

As an example of this formalism, we can declare an object that represents medication information and then populate it by querying the database.

med := OBJECT [code, route];
pt_meds := READ AS med (code,
routeCode) from substanceAdministration
where classCode = 'SBADM'
and moodCode = 'EVN';

In this example, *pt_meds* would be a list of *med* objects, and *first pt_meds.code* would be the vocabulary code of the first *med* object in the list returned by the query.

Finally, to avoid having to redefine operators to accommodate objects as operands, we require that operands be referenced in such a way using dot notation that the value yielded is a current Arden primitive data type.

Use of the New Model

Given a standard data model, query syntax and a format for referencing standard vocabularies workers at an institution can map their local vocabularies and database schema to the standard once, with some maintenance as the RIM or the vocabularies change. Then, the process of translation can occur automatically. Similarly, those wishing to share computable knowledge widely will write MLMs using these standards in order to make the process of using them at different institutions far easier than is currently the case.

DISCUSSION

While the proposed model should ease the process of adapting computable knowledge written at one institution for use at another, it does not eliminate all of this work. In order to use MLMs written using standard database linkages, an institution still must map its local vocabulary and schema to the standard. Moreover, that mapping must be maintained as both the local environment and the standard changes over time. However, once this is accomplished, translation of database linkages in MLMs can be automated, thus improving the likelihood of knowledge sharing.

Nevertheless, some adaptation beyond database references typically is required. In some cases, knowledge in a KB must be refined to meet local conditions, such as availability of certain kinds of diagnostic tests or therapies. Thus, adaptation of the logic of an MLM still may be required, even if the database references are standardized.

Although the challenge of adapting database references is mentioned commonly with regard to the Arden Syntax, this is true because Arden is the only standard for computable, procedural knowledge. As a result, this challenge has come to be called the "curly braces problem." However, any knowledge representation formalism that purports to facilitate knowledge sharing must address these issues. Thus, this challenge is not unique to Arden.

Finally, although others have specified an expression language and syntactic constructs for providing a standard data model and handling the resultant object variables, a concern with this approach is that it is not backward-compatible. Thus, compliance with such a standard would invalidate significant investment in current Arden KBs. By contrast, our approach provides the advantages of a standard data model and some features of an object-oriented environment while preserving backward-compatibility. This may serve as a transition to a full-fledged object-oriented model in the future.

FUTURE WORK

In concert with the Clinical Decision Support Technical Committee of HL7 of which one of us (RAJ) is co-chair, we will work to continue to refine this model and include it in the next version (2.5) of Arden Syntax. This refinement will include defining the behavior of aggregation operators when their operands are objects and how to handle the primary time of an object. Further, the principles of that underlie a standard data model must be applied to trigger event and alert destination mappings. Finally, we will ascertain the utility of including a flag in a database query that signals whether the query should be executed using an exact vocabulary code as written

or whether that code should be expanded at run time to include descendants of the vocabulary concept (query by class).

SUMMARY

In order to facilitate sharing of computable medical knowledge, we propose extensions to the Arden Syntax by defining a standard way to encode database linkages and to manipulate complex objects. By standardizing database linkages, we ease the process of automating knowledge adaptation when units of knowledge are shared between institutions.

Acknowledgments

We gratefully acknowledge the support of the Agency for Healthcare Research and Quality, grant R01-HS10472-01A1.

References

1. Jenders RA, Dasgupta B. Challenges in implementing a knowledge editor for the Arden Syntax: knowledge base maintenance and standardization of database linkages. Proc AMIA Symp 2001;:355-359.

2. Cabana MD, Rand CS, Powe NR et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999;282:1458-1465.

3. Bates DW, Cohen M, Leape LL et al. Reducing the frequency of errors in medicine using information technology. JAMIA 2001;8:299-308.

4. Shiffman RN, Liaw Y, Brandt CA et al. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. JAMIA 1999;6:104-114.

5. Jenders RA, Hripcsak G, Sideli RV et al. Medical decision support: experience with implementing the Arden Syntax at the Columbia-Presbyterian Medical Center. Proc AMIA Symp 1995;:169-73.

6. Pryor TA, Hripcsak G. Sharing MLM's: an experiment between Columbia-Presbyterian and LDS Hospital. Proc AMIA Symp 1993;:399–403.

7. Sujansky W, Altman R. Towards a standard query model for sharing decision-support applications. Proc AMIA Symp 1994;:325-331.

8. Jenders RA, Sujansky W, Broverman C, Chadwick M. Toward improved knowledge sharing: assessment of the HL7 Reference Information Model to support medical logic module queries. Proc AMIA Symp 1997;:308-312.

9. Hripcsak G, Ludemann P, Pryor TA, Wigertz OB, Clayton PD. Rationale for the Arden Syntax. Comput Biomed Res 1994;27(4):291-324.

10. Peleg M, Ogunyemi O, Tu S et al. Using features of Arden Syntax with object-oriented data models for guideline modeling. Proc AMIA Symp 2001;:523-527.

11. Peleg M, Boxwala AA, Bernstam E, Tu S, Greenes RA, Shortliffe EH. Sharable representation of clinical guidelines in GLIF: relationship to the Arden Syntax. J Biomed Inform 2001;34(3):170-181.

12. Huff SM, Rocha RA, McDonald CJ et al. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. JAMIA 1998;5(3):276-92.

13. Wang AY, Sable JH, Spackman KA. The SNOMED Clinical Terms development process: refinement and analysis of content. Proc AMIA Symp 2002;:845-849.

14. Cimino JJ, Clayton PD, Hripcsak G et al. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA 1994;1(1):35-50.